

Dynamic Environments for Virtual Machine Placement considering Elasticity and Overbooking

Jammily Ortigoza
Science and Technology School
Catholic University of Asunción
jortigozaf@gmail.com
Paraguay

Fabio López-Pires
Itaipu Technological Park
National University of Asunción
fabio.lopez@pti.org.py
Paraguay

Benjamín Barán
National University of Asunción
Catholic University of Asunción
bbaran@pol.una.py
Paraguay

Abstract—Cloud computing datacenters provide millions of virtual machines in actual cloud markets. In this context, Virtual Machine Placement (VMP) is one of the most challenging problems in cloud infrastructure management, considering the large number of possible optimization criteria and different formulations that could be studied. Considering the on-demand model of cloud computing, the VMP problem should be solved dynamically to efficiently attend typical workload of modern applications. This work proposes a taxonomy in order to understand possible challenges for Cloud Service Providers (CSPs) in dynamic environments, based on the most relevant dynamic parameters studied so far in the VMP literature. Based on the proposed taxonomy, several unexplored environments have been identified. To further study those research opportunities, sample workload traces for each particular environment are required; therefore, basic examples illustrate a preliminary work on dynamic workload trace generation.

I. INTRODUCTION

The rapid demand growth for computational resources in modern business and scientific applications presents several challenges for design, implementation and management of scalable datacenters to meet the requirements of customers in a competitive and efficient way [67].

Considering the evolution of resource provisioning, three main models could be identified: (1) traditional provisioning of resources with independent physical hardware, (2) modern provisioning of shared resources through virtualized hardware and (3) trending dynamic provisioning of resources through a cloud computing model [6]. The traditional provisioning environment has mostly evolved to a virtualized provisioning of resources in current datacenters, considering its advantages for management and resource utilization.

Virtualization in modern datacenters introduces complex management decisions related to the placement of virtual machines (VMs) into the available physical machines (PMs). In this context, Virtual Machine Placement (VMP) represents the process of selecting which VMs should be executed in a given set of PMs of a datacenter [47]. The VMP problem is mostly formulated as a combinatorial optimization problem, representing one of the most challenging problems in virtualized datacenters infrastructure management, considering the large number of possible optimization criteria and different formulations that could be studied [49].

For virtualized datacenters with deployments of VMs that rarely change its configuration over time, a static (offline) formulation of the VMP problem may be appropriate [50]. Additionally, in virtualized datacenters where a small number of VMs are created and destroyed, a semi-static formulation of the VMP problem could be acceptable (e.g. consolidating VMs every day at midnight). On the other hand, considering the today more realistic on-demand model of cloud computing with dynamic resource provisioning, a static (or semi-static) formulation of the VMP problem can result in under-optimal solutions after a short period of time. Clearly, the VMP problem for cloud computing environments must be formulated as a pure dynamic (online) optimization problem to efficiently attend dynamic workload of modern applications [49].

A. Background and Motivation

The VMP problem has been extensively studied and several surveys have already been presented in the VMP literature. Existing surveys focus on specific issues such as: (1) energy-efficient techniques applied to the problem [4], [60], (2) particular architectures where the VMP problem is applied, as federated clouds [22], and (3) methods for comparing performance of placement algorithms in large on-demand clouds [52]. None of the mentioned surveys presented a general and extensive study of a large part of the VMP literature. In consequence, López-Pires and Barán presented in [49] an extensive up-to-date survey of the most relevant VMP literature and proposed a novel taxonomy in order to identify research opportunities defining a general vision on this problem.

According to [49], the VMP problem is mostly formulated as an online optimization problem, where live migration techniques allow VMs to be dynamically consolidated on necessary PMs according to dynamic requirements of resources. The most studied environment for online formulations of the VMP problem considers that VMs are dynamically created and destroyed [49]. To the best of the authors' knowledge, there is no published work presenting a detailed characterization of possible dynamic environments for the VMP problem.

Clearly, a deeper research of possible dynamic parameters in cloud computing is necessary in order to propose holistic and more realistic environments for the formulation of the VMP problem for cloud computing datacenters.

Consequently, this work complements the taxonomy presented by the authors in [49] focusing specifically on dynamic formulations of the VMP problem from the providers' perspective, proposing a taxonomy in order to understand possible challenges for Cloud Service Providers (CSPs) in dynamic environments to efficiently attend customers' requests for virtual resources, based on the most relevant dynamic parameters studied so far in the VMP literature. The taxonomy proposed in this work must be jointly studied with the taxonomy first proposed in [49] in order to represent a complete VMP problem.

The remainder of this paper is organized as follows: Section II details the literature selection process considered in this work, while Section III introduces the classification criteria of the proposed taxonomy. Section IV presents the proposed taxonomy detailing the mathematical notation and basic examples of the identified dynamic environments for the VMP problem. Based on the proposed taxonomy, Section V presents a preliminary work on generation of workload traces for the identified dynamic environments. Finally, conclusions and future work are left to Section VI.

II. REVIEWED LITERATURE

A. Keywords Search

The selection process of relevant articles started with a search for research articles from Google Scholar database [scholar.google.com] with at least one of the following selected keywords in the article title: (1) virtual machine placement, (2) vm placement, (3) virtual machine consolidation, (4) vm consolidation or (5) server consolidation.

B. Publisher Filtering

Considering the large number of results from keywords search step, the literature selection process focused on research articles from the following well-known publishers: (1) ACM, (2) IEEE, (3) Elsevier and (4) Springer. This filtering step resulted in a reduction from 446 to 172 research articles. A detailed list of the 172 resulting articles can be found in [48].

C. Abstract Reading

Considering the 172 resulting articles from the publisher filtering step, a reading of the abstracts was performed in order to identify the most relevant articles that specifically study the VMP problem. Additionally, short papers (i.e. research articles with less than 6 pages) were removed from the selected literature, resulting in 84 selected articles of the VMP literature. A detailed list of the 84 resulting articles can be found in [48].

D. Online Formulations for Provider-oriented VMP Problem

Based on the 84 studied articles addressed in [49], this work selected the 64 articles that proposed online formulations for the VMP problem from the providers' perspective, considering the relevance of this type of environments for actual cloud computing providers. An in-depth reading of this universe of 64 articles was performed with the aim of identifying the most relevant dynamic parameters.

III. CLASSIFICATION CRITERIA

This work identified the following dynamic parameters:

- resource capacities of VMs (vertical elasticity);
- number of VMs of a service (horizontal elasticity);
- utilization of resources of VMs (related to overbooking).

Consequently, dynamic environments for online formulations of the provider-oriented VMP problem may be classified by one or more of the following classification criteria: (1) elasticity and (2) overbooking.

A. Elasticity

Considering the dynamic workload of modern cloud applications, proactive elasticity is a very important issue to address for CSPs in order to deal with under-provisioning (saturation) and over-provisioning (under-utilization) of cloud resources [3]. Under-provisioning can cause SLA violations, impacting directly on economical revenue while over-provisioning can cause inefficient utilization of resources, directly impacting on resource utilization and energy consumption.

Research articles considering online formulations of the provider-oriented VMP literature have already studied two types of elasticity: vertical and horizontal (see Figure 1). Vertical elasticity can be defined as the ability of cloud services to dynamically change capacities of virtual resources (e.g. CPU and RAM memory) inside a VM, while horizontal elasticity can be defined as the ability of cloud services to dynamically adjust the number of VMs [73].

It should be noted that from a CSP perspective, cloud services considering elasticity should be more important (i.e. higher level of SLA) than other non-elastic cloud services. An elastic cloud application could request additional resources to scale up the applications' resources and a CSP must consider more important the mentioned request than a request from a not elastic cloud service.

Implementing vertical elasticity requires shorter time of service reconfiguration than horizontal elasticity, but with a higher migration cost. On the other hand, horizontal elasticity enables stronger high availability than vertical elasticity, but a coordination overhead is required and infrastructure complexity increases [73].

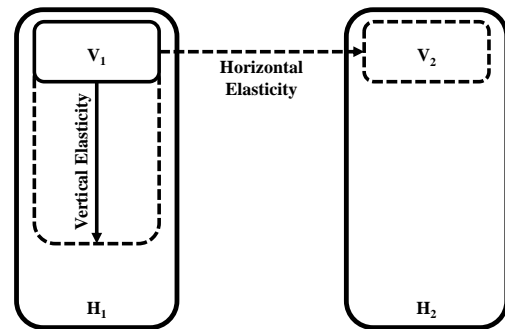


Figure 1. Vertical and horizontal elasticity. Vertical elasticity dynamically adjusts the capacities of virtual resources inside a VM while horizontal elasticity dynamically adjusts the number of VMs (e.g. in distributed applications).

B. Overbooking

Resources overbooking can make cloud services more profitable for CSPs, overlaying requested virtual resources onto physical resources at a higher ratio than 1:1 [29]. Online formulations of the provider-oriented VMP considering overbooking include particular considerations to efficiently attend customers' requirements, enforcing SLAs.

Considering the dynamic workload of cloud applications and services, virtual resources of VMs are also dynamically used giving space to re-utilization of idle resources that were already reserved. Research articles considering online formulations of the provider-oriented VMP literature have already studied two types of overbooking: server and network resources overbooking.

In this context, CSPs should measure the utilization of resources of VMs in order to correctly manage the overbooking with the available physical resources, minimizing SLA violations. Monitoring utilization of virtual resources and workload of cloud applications and services also helps CSPs to consider forecasting techniques for approximating in advance the required management actions (e.g. migrations of VMs) for the consolidation process, to reduce resource under-provisioning [19], [77].

IV. PROPOSED TAXONOMY

Based on the universe of 64 studied research articles, dynamic environments for online formulations of the provider-oriented VMP problem may be classified by one or more of the following classification criteria: (1) elasticity and (2) overbooking, as presented in Section III.

The proposed taxonomy is presented in Figure 3 as a two-dimensional coordinate axis where each dimension represents a classification criteria (elasticity and overbooking).

First, dynamic environments could be formulated considering one of the following elasticity values:

- elasticity=0: no elasticity;
- elasticity=1: horizontal elasticity;
- elasticity=2: vertical elasticity;
- elasticity=3: horizontal and vertical elasticity.

Additionally, identified dynamic environments may also consider one of the following overbooking values:

- overbooking=0: no overbooking;
- overbooking=1: server resources overbooking;
- overbooking=2: network resources overbooking;
- overbooking=3: server and network overbooking.

Based on the combinations of the possible values of the classification criteria (elasticity, overbooking), the proposed taxonomy identified 16 different possible environments (see Figure 3). Considering this representation, each identified dynamic environment is denoted by its elasticity and overbooking coordinates. For example, Environment (0,0) denotes a dynamic environment that does not consider neither any type of elasticity nor any type of overbooking, while Environment (1,3) denotes a dynamic environment that considers horizontal elasticity with both types of overbooking.

It should be mentioned that all the identified environments in this work consider that VMs are dynamically created and destroyed. According to the identified dynamic environments, the simplest environment is Environment (0,0), while the most complex environment is Environment (3,3). Additionally, the proposed taxonomy showed that 50% of the articles studied Environment (0,1) while 39% of the studied articles considered Environment (0,0), representing the most studied environments in the considered literature. Several unexplored environments were also identified, as detailed in the following subsections.

A. Cloud Service and Environment Notation

CSPs dynamically receive requests for the placement of cloud services with different characteristics according to the classification criteria presented in Section III, representing real-world environments and generalizing the deployment of cloud services in several possible cloud architectures (e.g. single-cloud, distributed-cloud or federated-cloud). Cloud services may represent simple services such as Domain Name Service (DNS) or complex multi-tier elastic applications.

A cloud service is composed by a set of VMs, where each VM of a cloud service could be located for its execution in different cloud datacenters according to the customers preferences or requirements (e.g. legal issues or high-availability).

Configuration of VMs of a cloud service changes dynamically when elasticity is considered. On the other hand, utilization of virtual resources change dynamically according to the demand when overbooking is considered; otherwise, the utilization of each virtual resource is considered at 100%.

Formally, a cloud service S_b can be distributed across different possible cloud datacenters. Each cloud datacenter DC_c hosts VMs V'_{cj} associated to different cloud services. A VM V'_{cj} associated to a service S_b is denoted as V''_{bcj} .

where:

S_b :	Cloud service b ;
DC_c :	Cloud datacenter c ;
mDC_c :	Number of VMs V_j in DC_c ;
mS_b :	Number of VMs V_j in S_b ;
V'_{cj} :	V_j in DC_c ;
V''_{bcj} :	V_j in DC_c from service S_b .

Figure 2 presents a basic example of a cloud service S_1 , distributed across 2 cloud datacenters DC_1 and DC_2 and using 4 VMs: V''_{111} , V''_{112} , V''_{121} , V''_{122} . These cloud datacenters could represent geo-distributed datacenters owned by one CSP or a federated-cloud with two different CSPs. Each cloud datacenter hosts 2 VMs of S_1 : V'_{11} and V'_{12} represent V_1 and V_2 in DC_1 respectively (analogously DC_2 hosts 2 VMs).

Complementing the above notation, each cloud datacenter DC_c may be represented as:

$$DC_c = \{V'_{cj}, j \in \{1, 2, \dots, mDC_c\}\} \quad (1)$$

For simplicity, this work considers only processing, memory and network resources for a VM, but the notation is general enough for considering any set of virtual resources.

$$V''_{bcj} = \{Vcpu''_{bcj}, Vram''_{bcj}, Vnet''_{bcj}, R''_{bcj}, SLA''_{bcj}, t_{init}, t_{end}\} \quad (2)$$

where:

V''_{bcj} : V_j in DC_c from service S_b ;
 $Vcpu''_{bcj}$: Processing requirements of V''_{bcj} in [ECU];
 $Vram''_{bcj}$: Memory requirements of V''_{bcj} in [GB];
 $Vnet''_{bcj}$: Network requirements of V''_{bcj} in [Mbps];
 R''_{bcj} : Economical revenue for locating V''_{bcj} in [\$];
 SLA''_{bcj} : SLA of V''_{bcj} . $SLA''_{bcj} \in \{1, \dots, s\}$;
 s : Highest priority level of SLAs;
 t_{init} : Initial discrete time when V''_{bcj} is executed;
 t_{end} : Final discrete time when V''_{bcj} is executed.

Utilization of the resources of each V''_{bcj} is represented by:

$$U''_{bcj} = \{Ucpu''_{bcj}, Uram''_{bcj}, Unet''_{bcj}\} \quad (3)$$

where:

U''_{bcj} : Utilization of requirement V''_{bcj} ;
 $Ucpu''_{bcj}$: Utilization of $Vcpu''_{bcj}$ in [ECU];
 $Uram''_{bcj}$: Utilization of $Vram''_{bcj}$ in [GB];
 $Unet''_{bcj}$: Utilization of $Vnet''_{bcj}$ in [Mbps].

Note that in practical applications $Ucpu''_{bcj}$ is lower than $Vcpu''_{bcj}$ ($Ucpu''_{bcj} \leq Vcpu''_{bcj}$), giving place to CPU overbooking of resources. The same overbooking situation may occur for other resources.

Each of the 16 identified environments (see Figure 3) considers different parameters that dynamically change as a function of time t , giving place to possible different notations for each environment. The following subsections detail each of the identified dynamic environment, presenting particular notation for its characterization.

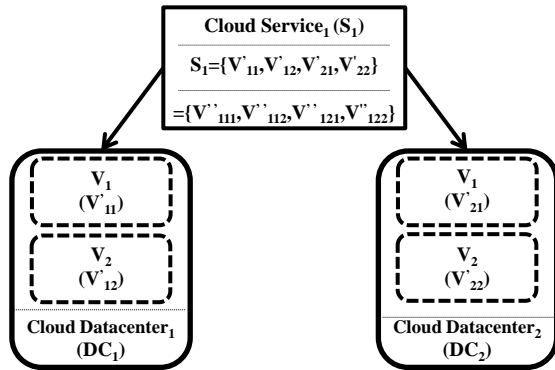


Figure 2. Example of a cloud service considered in this work.

B. Dynamic Environment Classification

Based on the notation previously presented in Section IV-A, the 16 identified environments (see Figure 3) present particular considerations and different time variables may be defined for a correct characterization. A summary of the time variables is presented in Table I. Next, the identified dynamic environments are presented, enumerated by its elasticity and overbooking coordinates.

- **(0,0) No Elasticity (0), No Overbooking (0)**: It represents the most basic dynamic environment identified for solving the provider-oriented VMP problem. The CSP have to attend the placement of cloud services that are dynamically created and destroyed in function of time t . From the studied universe of 64 articles, 39% proposed formulations of the VMP problem considering this basic environment (Figure 3).
- **(0,1) No Elasticity (0), Server Resources Overbooking (1)**: According to the studied articles, the provider-oriented VMP problem is mostly formulated considering overbooking of server resources (i.e. processing, memory and storage) without considering neither horizontal nor vertical elasticity. This environment represents 50% of the studied universe of 64 articles (see Figure 3). For this particular environment, CSPs must monitor the dynamic utilization of virtual server resources for a safe overbooking. The following variables must be defined as a function of time: $Ucpu''_{bcj}(t)$ and $Uram''_{bcj}(t)$.
- **(0,2) No Elasticity (0), Network Resources Overbooking (2)**: Overbooking could be also considered exclusively for virtual network resources. Analogously to the Environment (0,1), CSPs must monitor the dynamic utilization of virtual network resources for a safe overbooking. Consequently, the utilization of network resources is defined as a time variable: $Unet''_{bcj}(t)$. This environment represents only 5% of the studied articles (see Figure 3).
- **(0,3) No Elasticity (0), Server and Network Resources Overbooking (3)**: Representing the most complex environment for overbooking, this dynamic environment is identified as a research opportunity, considering that no studied article proposed a formulation of the provider-oriented VMP problem that jointly considers overbooking of server and network resources without elasticity (see Figure 3). For this particular environment, CSPs must monitor the dynamic utilization of both virtual server and network resources for a safe overbooking. Consequently, $Ucpu''_{bcj}(t)$, $Uram''_{bcj}(t)$ and $Unet''_{bcj}(t)$ are defined as a function of time.
- **(1,0) Horizontal Elasticity (1), No Overbooking (0)**: Elasticity could be considered in order to efficiently attend the dynamic demand of resources according to a SLA associated to a given cloud service. A dynamic environment that considers horizontal elasticity represents particular considerations associated to scaling up and down the number of requested VMs that composes a cloud service.

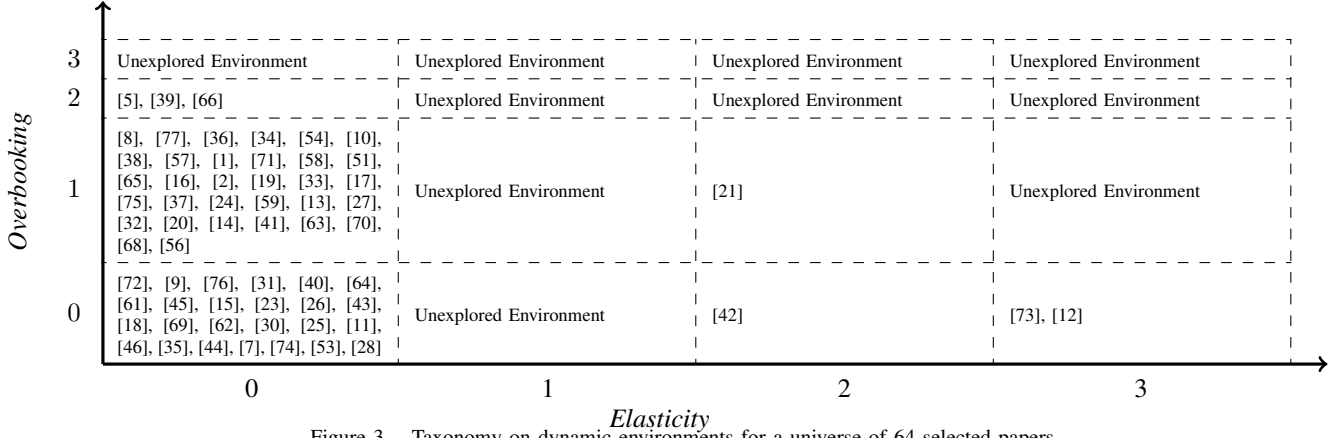


Figure 3. Taxonomy on dynamic environments for a universe of 64 selected papers.

Determining when to scale is not considered a responsibility of the CSPs and it is out of the scope of this work. In this context, the number of VMs of cloud services is a time variable: $mS_{b_{min}} \leq mS_b(t) \leq mS_{b_{max}}$. This particular environment is also identified as a research opportunity, representing the most basic environment for solving problems considering horizontal elasticity for parallel applications such as MapReduce jobs.

- **(1,1) Horizontal Elasticity (1), Server Resources Overbooking (1):** Additionally to horizontal elasticity, a dynamic environment could also include overbooking of server resources. The following variables are defined as a function of time: $Ucpu''_{bcj}(t)$, $Uram''_{bcj}(t)$ and $mS_b(t)$. This environment also represents a research opportunity.
- **(1,2) Horizontal Elasticity (1), Network Resources Overbooking (2):** Analogously to the Environment (1,1), CSPs must monitor the dynamic utilization of virtual network resources for a safe overbooking. Consequently, $Unet''_{bcj}(t)$ is defined as a time variable, additionally to $mS_b(t)$. This environment is identified as a research opportunity, considering that no studied article considers overbooking of network resources with horizontal elasticity.
- **(1,3) Horizontal Elasticity (1), Server and Network Resources Overbooking (3):** No studied article proposed a formulation of the provider-oriented VMP problem that jointly considers overbooking of server and network resources with horizontal elasticity, representing a research opportunity. For this particular environment, the following variables are defined as a function of time: $Ucpu''_{bcj}(t)$, $Uram''_{bcj}(t)$, $Unet''_{bcj}(t)$ and $mS_b(t)$.
- **(2,0) Vertical Elasticity (2), No Overbooking (0):** As mentioned before, elasticity could be considered in order to efficiently attend the dynamic demand of resources according to a SLA associated to a cloud service. A dynamic environment that considers vertical elasticity represents particular considerations associated to the virtual resources capacities of requested VMs that composes a cloud service. This work considers processing and memory requirements as time variables: $Vcpu''_{bcj}(t)$ and $Vram''_{bcj}(t)$.

It should be mentioned that the notation presented in this section is general enough to consider any other resources for vertical elasticity such as $Vnet''_{bcj}(t)$ just to cite one. According to the studied articles, only [42] (1.5%) studied this environment (see Figure 3).

- **(2,1) Vertical Elasticity (2), Server Resources Overbooking (1):** Additionally to vertical elasticity, a dynamic environment could also include overbooking of server resources. For these particular environment, CSPs must monitor the dynamic utilization of virtual server resources for a safe overbooking. Consequently, the following variables are defined as a function of time: $Ucpu''_{bcj}(t)$, $Uram''_{bcj}(t)$, $Vcpu''_{bcj}(t)$ and $Vram''_{bcj}(t)$. According to the studied articles, only [21] (1.5%) studied this environment (see Figure 3).
- **(2,2) Vertical Elasticity (2), Network Resources Overbooking (2):** Analogously to the Environment (2,1), $Unet''_{bcj}(t)$ is defined as a time variable, additionally to $Vcpu''_{bcj}(t)$ and $Vram''_{bcj}(t)$. This particular environment represents a research opportunity, considering that no studied article proposed a formulation of the provider-oriented VMP in this particular environment.
- **(2,3) Vertical Elasticity (2), Server and Network Resources Overbooking (3):** No studied article proposed a formulation of the provider-oriented VMP problem that jointly considers overbooking of server and network resources with vertical elasticity. For this particular environment, the following variables are defined as a function of time: $Ucpu''_{bcj}(t)$, $Uram''_{bcj}(t)$, $Unet''_{bcj}(t)$, $Vcpu''_{bcj}(t)$ and $Vram''_{bcj}(t)$.
- **(3,0) Horizontal and Vertical Elasticity (3), No Overbooking (0):** Both types of elasticity lead to different impacts for cloud datacenters infrastructure management and respond to different requirements of customers' applications. Definitely, in real world environments CSPs should be able to solve the VMP problem considering formulations that jointly implement both horizontal and vertical elasticity for cloud services. In this context, [73] and [12] proposed different approaches for dealing with these challenges, representing 3% of the studied universe (see Figure 3).

In this environment of both mixed elasticity types, the following time variables are defined: $mS_b(t)$, $Vcpu''_{bcj}(t)$ and $Vram''_{bcj}(t)$.

- **(3,1) Horizontal and Vertical Elasticity (3), Server Resources Overbooking (1):** Additionally to horizontal and vertical elasticity, a dynamic environment could also include overbooking of server resources. For these particular environment, CSPs must monitor the dynamic utilization of virtual server resources for a safe overbooking. Consequently, the following variables are defined as a function of time: $Ucpu''_{bcj}(t)$, $Uram''_{bcj}(t)$, $mS_b(t)$, $Vcpu''_{bcj}(t)$ and $Vram''_{bcj}(t)$.
- **(3,2) Horizontal and Vertical Elasticity (3), Network Resources Overbooking (2):** Analogously to the Environment (3,1), the following variables are defined in function of time: $mS_b(t)$, $Vcpu''_{bcj}(t)$, $Vram''_{bcj}(t)$ and $Unet''_{bcj}(t)$. This particular environment represent a research opportunity, considering that no studied article proposed a formulation of the provider-oriented VMP in this particular environment.
- **(3,3) Horizontal and Vertical Elasticity (3), Server and Network Resources Overbooking (3):** Considering both types of elasticity and both types of overbooking represent the most complex environment identified in this work. CSPs efficiently solving formulations of the VMP problem in this complex dynamic environment will represent a considerable advance on this research area and its cloud datacenters will be able to scale according to trending types of requirements with sufficient flexibility. As the most general environment, the following variables are defined in function of time for characterizing this complex environment: $mS_b(t)$, $Vcpu''_{bcj}(t)$, $Vram''_{bcj}(t)$, $Ucpu''_{bcj}(t)$, $Uram''_{bcj}(t)$ and $Unet''_{bcj}(t)$. A recommended path for research is exploring and addressing challenges of particular environments identified as research opportunities before considering this advanced and complete dynamic environment for solving the provider-oriented VMP problem.

C. Dynamic Environment Examples

Due to space limitations, this work focus on a representative example of the Environment (1,0), representing an elastic application implementing horizontal elasticity (see Figure 4). Interested readers can refer to [55] for a complete set of examples of the 16 identified dynamic environments.

It is important to remember that Environment (1,0) includes only horizontal elasticity, but vertical elasticity as well as both types of overbooking (server and network resources) can be observed in the workload trace presented in Section V-A.

Figure 4 presents different levels of detail for the environment. The CSP level (CSP_1) represents the requests that CSPs receive for the placement of cloud services (or VMs) in the PMs of the available cloud datacenters. Next, cloud service level (S_1) details requested resources of cloud services at each discrete time. Cloud datacenter levels (DC_1 and DC_2) detail resources of cloud services for each cloud datacenter.

Table I
SUMMARY OF TIME VARIABLES FOR THE 16 IDENTIFIED DYNAMIC ENVIRONMENTS.

Env.	Elasticity Type	Overbooking Type	Time Variables
(0,0)	Not Considered	Not Considered	-
(0,1)	Not Considered	Server	$-Ucpu''_{bcj}(t)$ $-Uram''_{bcj}(t)$
(0,2)	Not Considered	Network	$-Unet''_{bcj}(t)$
(0,3)	Not Considered	Server and Network	$-Ucpu''_{bcj}(t)$ $-Uram''_{bcj}(t)$ $-Unet''_{bcj}(t)$
(1,0)	Horizontal	Not Considered	$-mS_b(t)$
(1,1)	Horizontal	Server	$-mS_b(t)$ $-Ucpu''_{bcj}(t)$ $-Uram''_{bcj}(t)$
(1,2)	Horizontal	Network	$-mS_b(t)$ $-Unet''_{bcj}(t)$
(1,3)	Horizontal	Server and Network	$-mS_b(t)$ $-Ucpu''_{bcj}(t)$ $-Uram''_{bcj}(t)$ $-Unet''_{bcj}(t)$
(2,0)	Vertical	Not Considered	$-Vcpu''_{bcj}(t)$ $-Vram''_{bcj}(t)$
(2,1)	Vertical	Server	$-Vcpu''_{bcj}(t)$ $-Vram''_{bcj}(t)$ $-Ucpu''_{bcj}(t)$ $-Uram''_{bcj}(t)$
(2,2)	Vertical	Network	$-Vcpu''_{bcj}(t)$ $-Vram''_{bcj}(t)$ $-Unet''_{bcj}(t)$
(2,3)	Vertical	Server and Network	$-Vcpu''_{bcj}(t)$ $-Vram''_{bcj}(t)$ $-Ucpu''_{bcj}(t)$ $-Uram''_{bcj}(t)$ $-Unet''_{bcj}(t)$
(3,0)	Horizontal and Vertical	Not Considered	$-mS_b(t)$ $-Vcpu''_{bcj}(t)$ $-Vram''_{bcj}(t)$
(3,1)	Horizontal and Vertical	Server	$-mS_b(t)$ $-Vcpu''_{bcj}(t)$ $-Vram''_{bcj}(t)$ $-Ucpu''_{bcj}(t)$ $-Uram''_{bcj}(t)$
(3,2)	Horizontal and Vertical	Network	$-mS_b(t)$ $-Vcpu''_{bcj}(t)$ $-Vram''_{bcj}(t)$ $-Unet''_{bcj}(t)$
(3,3)	Horizontal and Vertical	Server and Network	$-mS_b(t)$ $-Vcpu''_{bcj}(t)$ $-Vram''_{bcj}(t)$ $-Ucpu''_{bcj}(t)$ $-Uram''_{bcj}(t)$ $-Unet''_{bcj}(t)$

The horizontal elasticity of the example in Figure 4 can be observed considering that initially, cloud service S_1 starts in $t = 0$ requesting 2 VMs (in blue) across cloud datacenters DC_1 and DC_2 from $t = 0$ to $t = 5$. Assuming an increasing demand for resources, S_1 scales up the number of VMs adding 1 VM hosted at cloud datacenter DC_1 (in brown) at $t = 1$. In $t = 2$, S_1 scales up the number of VMs adding 1 more VM hosted at cloud datacenter DC_2 (in brown) resulting in 4 VMs for attending the demand for resources from $t = 2$ to $t = 3$.

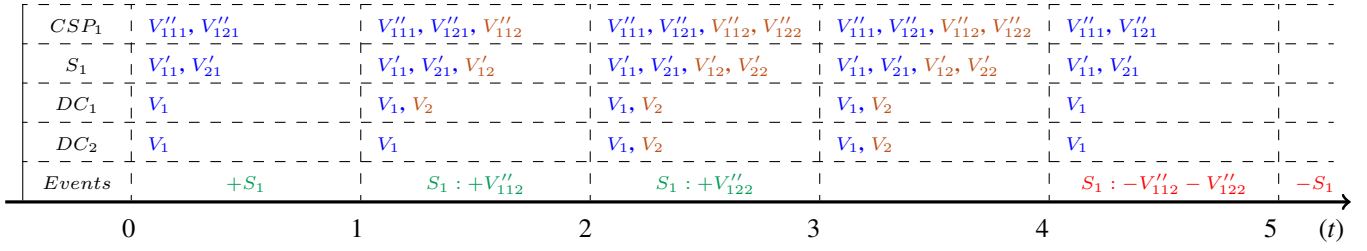


Figure 4. Basic example of Environment (1,0)

The cloud service S_1 returns to its initial configuration scaling down to 2 VMs at $t = 4$, assuming a decreasing demand for resources, finishing the requests for resources at $t = 5$.

V. WORKLOAD TRACES FOR DYNAMIC ENVIRONMENTS

Based on the proposed taxonomy, several research opportunities have been identified (see Section IV-B). Therefore, sample workload traces for each particular environment are required in order to: (1) explore the challenges associated to each environment, (2) propose formulations and test algorithms that solve these challenges with different workload types and (3) effectively compare performance and quality of different algorithms with reproducible experiments.

As identified in [48], there is no existing testbed problem instances for the VMP that can today be used as a world accepted benchmark. Consequently, the authors are working on a workload trace generator for the VMP problem to be able to generate different instances for experimental tests based on the dynamic environments proposed in this work. A brief introduction of preliminary results is presented in this section.

The proposed workload trace generator for the VMP problem considers the following input data (see Table II):

- workload trace duration,
- range of values for virtual resources of VMs,
- range of values for utilization of virtual resources of VMs,
- range of revenue values for executing VMs,
- range of SLA values of VMs,
- range of number of VMs for cloud services,
- number of cloud services,
- probability distribution.

Table II
INPUT DATA FOR EXAMPLE WORKLOAD TRACE FROM TABLE III.

Input Data	Value (Min - Max)
Workload trace duration (t)	(4 - 4)
Range of CPU values for resources	(4 - 10)
Range of Memory values for resources	(2 - 16)
Range of Network values for resources	(100 - 1000)
Range of CPU values for utilization	(2 - 10)
Range of Memory values for utilization	(1 - 16)
Range of Network values for utilization	(0 - 1000)
Range of revenue values for executing VMs	(0.1 - 1.5)
Range of SLA values of VMs	(0 - 2)
Range of number of VMs for cloud services	(1 - 6)
Number of cloud services	(1 - 1)
Probability distribution	Random

It is important to mention that additionally, an user of the generator is also able to include real-world workload traces, extending or reducing the trace to specific requirements for the experiments. In this case, the workload generator could generate synthetic workloads based on real-world workloads adjusting the workload to a specific size of the problem instance (e.g. workload trace duration), maintaining characteristics of the real-world trace (e.g. probability distribution).

A. Workload Trace Generation Example

As an example of utilization of the workload trace generator, Table II presents the input data considered for the generation of the workload trace presented in Table III. Due to space limitation as well as similarity of the structure of the workload traces of the different environments, Table III represents only a basic example of the most complex dynamic environment identified in this work, Environment (3,3). This example includes all possible dynamic parameters (resource capacities of VMs, number of VMs of a cloud service and utilization of resources of VMs). Interested readers can refer to [55] for more detailed examples of workload traces for the 16 different dynamic environments.

According to Table II, all the considered values are selected randomly from the specified values considering that for this particular example the probability distribution is uniform. It is important to mention that several other data probability distributions could be considered.

The generated workload trace of Table III could have only a duration of $t = 4$ considering that both minimum and maximum values were 4 (see Table II). For this particular example, the number of cloud services was fixed to 1 considering that both minimum and maximum values were 1 (see Table II).

Additionally, the number of VMs of the cloud services in the example of Table III can be adjusted from 1 to 6. The revenue for executing VMs can vary from 0.1\$ to 1.5\$ and its corresponding SLAs can vary from 0 to 2. Analogously, the values for each virtual resources of VMs as well as its utilization can vary from its specified values (see Table II).

Horizontal elasticity is considered in Table III in order to efficiently attend the increasing demand of resources scaling up the number of VMs of S_1 from 2 (at $t = 0$) to 3 (at $t = 1$) and from 3 (at $t = 1$) to 4 (at $t = 2$). The number of VMs scales down from 4 (at $t = 3$) to 2 (at $t = 4$), assuming a decreasing demand of resources. Additionally, vertical elasticity for processing and memory resources could

Table III
EXAMPLE OF WORKLOAD TRACE FOR VMP PROBLEM IN ENVIRONMENT (3,3)

t	S_b	D_c	V_j	$Vcpu''_{bcj}$	$Vram''_{bcj}$	$Vnet''_{bcj}$	R''_{bcj}	SLA''_{bcj}	$Ucpu''_{bcj}$	$Uram''_{bcj}$	$Unet''_{bcj}$
0	1	1	1	8	16	1000	0.5	1	8	14	150
0	1	2	1	8	16	1000	0.5	1	8	9	50
1	1	1	1	8	16	1000	0.5	1	7	10	160
1	1	2	1	8	16	1000	0.5	1	7	10	100
1	1	1	2	8	16	1000	0.5	1	7	7	70
2	1	1	1	8	16	1000	0.5	1	6	11	200
2	1	2	1	8	16	1000	0.5	1	6	11	150
2	1	1	2	8	16	1000	0.5	1	6	9	50
2	1	2	2	8	16	1000	0.5	1	6	12	60
3	1	1	1	8	16	1000	0.5	1	4	12	180
3	1	2	1	8	16	1000	0.5	1	4	12	150
3	1	1	2	8	16	1000	0.5	1	1	9	60
3	1	2	2	8	16	1000	0.5	1	1	8	60
4	1	1	1	4	8	1000	0.5	1	2	6	200
4	1	2	1	4	8	1000	0.5	1	2	6	100

be observed in the workload trace of Table III from $t = 3$ to $t = 4$, where $Vcpu''_{111}$ and $Vcpu''_{121}$ decrease from 8 [ECU] to 4 [ECU] and $Vram''_{111}$ and $Vram''_{121}$ decrease from 16 [GB] to 8 [GB], assuming a decreasing demand of resources. Vertical elasticity could also be applied to other resources as described in Section IV-B.

Finally, server and network resources utilization change dynamically in VMs from $t = 0$ to $t = 4$, representing important data for CSPs in order to apply a safe overbooking of both server and network resources (see Table III). At $t = 0$, it can be seen a high utilization of both processing and memory resource, representing a possible alarm for scaling up the number of VMs (horizontal elasticity) as can be observed in $t = 1$. Low utilization of resources can be seen at $t = 3$, representing an alarm for scaling down both the number of VMs (horizontal elasticity) as well as hardware configuration of each VM (vertical elasticity) as can be seen at $t = 4$.

VI. CONCLUSIONS AND FUTURE WORK

Based on an universe of 64 studied publications carefully chosen as explained in Section II, this work extended the taxonomy presented in [49] focusing on dynamic (online) formulations of the VMP problem from the providers' perspective complementing a previous work of the authors [49] and proposed a novel taxonomy in order to understand possible challenges for Cloud Service Providers (CSPs) in dynamic environments to efficiently attend customers' request for virtual resources, based on the most relevant dynamic parameters studied so far in the VMP literature.

This work identified that resource capacities of VMs (associated to vertical elasticity), number of VMs of a cloud service (associated to horizontal elasticity) and utilization of resources of VMs (related to overbooking) are the most relevant dynamic parameters in literature. Consequently, dynamic environments for online formulations of the provider-oriented VMP problem were classified by one of the following classification criteria: (1) elasticity and (2) overbooking. First, dynamic environments could be formulated considering one of the following

elasticity values: no elasticity, horizontal elasticity, vertical elasticity or both horizontal and vertical elasticity. Additionally, identified dynamic environments may also consider one of the following overbooking values: no overbooking, server resources overbooking, network resources overbooking or both server and network overbooking.

Based on the combinations of the possible values of the classification criteria (elasticity and overbooking), the proposed taxonomy identified 16 different possible environments (see Figure 3), characterizing each environment with particular mathematical notation of time variables (see Table I).

The proposed taxonomy showed that research of online formulations of the provider-oriented VMP problem has been mainly studied in Environment (0,0) and (0,1) with 39% and 50% of the studied articles respectively. Other briefly studied environments are Environment (0,2), (2,0), (2,1) and (3,0). Several research opportunities for unexplored environments were identified (see Figure 3). For example, no paper was found studying horizontal elasticity alone, even more, joint network and server overbooking is still a field with no published paper. Considering both types of elasticity and both types of overbooking represent the most advanced environment identified in this work: Environment (3,3). CSPs efficiently solving formulations of the VMP problem in this complex (3,3) dynamic environment will represent a considerable advance on this research area and its cloud datacenters will be able to scale according to trending types of requirements with sufficient flexibility. A recommended path for future work is exploring and addressing challenges of particular environments identified as research opportunities before considering this advanced and complete (3,3) dynamic environment for solving the provider-oriented VMP problem.

At the time of this writing, the authors are already working on extending identified environments to consider dynamic level of SLAs and dynamic revenue for executing VMs, just to cite a few characteristics to be included. Additionally, other environments could be studied considering dynamic electricity costs or pricing schemes in federated clouds, among others.

REFERENCES

- [1] M. Alicherry and T. Lakshman, "Optimizing data access latencies in cloud systems by intelligent virtual machine placement," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 647–655.
- [2] A. Anand, J. Lakshmi, and S. Nandy, "Virtual machine placement optimization supporting performance slas," in *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*, vol. 1. IEEE, 2013, pp. 298–305.
- [3] M. Armbrust, O. Fox, R. Griffith, A. D. Joseph, Y. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica *et al.*, "Above the clouds: A Berkeley view of cloud computing," *University of California, Berkeley, Tech. Rep.*, 2009.
- [4] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future Generation Computer Systems*, vol. 28, no. 5, pp. 755–768, 2012.
- [5] O. Biran, A. Corradi, M. Fanelli, L. Foschini, A. Nus, D. Raz, and E. Silvera, "A stable network-aware vm placement for cloud systems," in *Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012)*. IEEE Computer Society, 2012, pp. 498–506.
- [6] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities," in *High Performance Computing and Communications, 2008. HPCC'08. 10th IEEE International Conference on*, 2008.
- [7] N. M. Calcevachia, O. Biran, E. Hadad, and Y. Moatti, "Vm placement strategies for cloud scenarios," in *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*. IEEE, 2012, pp. 852–859.
- [8] Z. Cao and S. Dong, "An energy-aware heuristic framework for virtual machine consolidation in cloud computing," *The Journal of Supercomputing*, pp. 1–23, 2014.
- [9] D. Chang, G. Xu, L. Hu, and K. Yang, "A network-aware virtual machine placement algorithm in mobile cloud computing environment," in *Wireless Communications and Networking Conference Workshops (WCNCW), 2013 IEEE*. IEEE, 2013, pp. 117–122.
- [10] K.-y. Chen, Y. Xu, K. Xi, and H. J. Chao, "Intelligent virtual machine placement for cost efficiency in geo-distributed cloud systems," in *Communications (ICC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3498–3503.
- [11] A. Dalvandi, M. Gurusamy, and K. C. Chua, "Time-aware vm-placement and routing with bandwidth guarantees in green cloud data centers," in *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*, vol. 1. IEEE, 2013, pp. 212–217.
- [12] H. T. Dang and F. Hermenier, "Higher sla satisfaction in datacenters with continuous vm placement constraints," in *Proceedings of the 9th Workshop on Hot Topics in Dependable Systems*. ACM, 2013, p. 1.
- [13] D. S. Dias and L. H. M. Costa, "Online traffic-aware virtual machine placement in data center networks," in *Global Information Infrastructure and Networking Symposium (GIIS), 2012*. IEEE, 2012, pp. 1–8.
- [14] A. V. Do, J. Chen, C. Wang, Y. C. Lee, A. Y. Zomaya, and B. B. Zhou, "Profiling applications for virtual machine placement in clouds," in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*. IEEE, 2011, pp. 660–667.
- [15] D. Dong and J. Herbert, "Energy efficient vm placement supported by data analytic service," in *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on*. IEEE, 2013, pp. 648–655.
- [16] J. Dong, H. Wang, X. Jin, Y. Li, P. Zhang, and S. Cheng, "Virtual machine placement for improving energy efficiency and network performance in iaas cloud," in *Distributed Computing Systems Workshops (ICDCSW), 2013 IEEE 33rd International Conference on*. IEEE, 2013, pp. 238–243.
- [17] C. Dupont, G. Giuliani, F. Hermenier, T. Schulze, and A. Somov, "An energy aware framework for virtual machine placement in cloud federated data centres," in *Future Energy Systems: Where Energy, Computing and Communication Meet (e-Energy), 2012 Third International Conference on*. IEEE, 2012, pp. 1–10.
- [18] S. Fang, R. Kanagavelu, B.-S. Lee, C. H. Foh, and K. M. M. Aung, "Power-efficient virtual machine placement and migration in data centers," in *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCoM), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*. IEEE, 2013, pp. 1408–1413.
- [19] W. Fang, X. Liang, S. Li, L. Chiaraviglio, and N. Xiong, "Vmplanner: Optimizing virtual machine placement and traffic flow routing to reduce network power costs in cloud data centers," *Computer Networks*, vol. 57, no. 1, pp. 179–196, 2013.
- [20] T. Ferreto, C. A. De Rose, and H.-U. Heiss, "Maximum migration time guarantees in dynamic server consolidation for virtualized data centers," in *Euro-Par 2011 Parallel Processing*. Springer, 2011, pp. 443–454.
- [21] T. C. Ferreto, M. A. Netto, R. N. Calheiros, and C. A. De Rose, "Server consolidation with migration control for virtualized data centers," *Future Generation Computer Systems*, vol. 27, no. 8, pp. 1027–1034, 2011.
- [22] M. Gahlawat and P. Sharma, "Survey of virtual machine placement in federated clouds," in *Advance Computing Conference (IACC), 2014 IEEE International*. IEEE, 2014, pp. 735–738.
- [23] S. Georgiou, K. Tsakalozos, and A. Delis, "Exploiting network-topology awareness for vm placement in iaas clouds," in *Cloud and Green Computing (CGC), 2013 Third International Conference on*. IEEE, 2013, pp. 151–158.
- [24] H. Goudarzi and M. Pedram, "Energy-efficient virtual machine replication and placement in a cloud computing system," in *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*. IEEE, 2012, pp. 750–757.
- [25] Y. Guo, A. L. Stolyar, and A. Walid, "Shadow-routing based dynamic algorithms for virtual machine placement in a network cloud," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 620–628.
- [26] A. Gupta, L. V. Kalé, D. Milojicic, P. Faraboschi, and S. M. Balle, "Hpc-aware vm placement in infrastructure clouds," in *Cloud Engineering (IC2E), 2013 IEEE International Conference on*. IEEE, 2013, pp. 11–20.
- [27] A. Gupta, D. Milojicic, and L. V. Kalé, "Optimizing vm placement for hpc in the cloud," in *Proceedings of the 2012 workshop on Cloud services, federation, and the 8th open cirrus summit*. ACM, 2012, pp. 1–6.
- [28] Y. Ho, P. Liu, and J.-J. Wu, "Server consolidation algorithms with bounded migration cost and performance guarantees in cloud computing," in *Utility and Cloud Computing (UCC), 2011 Fourth IEEE International Conference on*. IEEE, 2011, pp. 154–161.
- [29] D. Hoefflin and P. Reeser, "Quantifying the performance impact of overbooking virtualized resources," in *Communications (ICC), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5523–5527.
- [30] H.-J. Hong, D.-Y. Chen, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu, "Qoe-aware virtual machine placement for cloud games," in *Network and Systems Support for Games (NetGames), 2013 12th Annual Workshop on*. IEEE, 2013, pp. 1–2.
- [31] W. Huang, X. Li, and Z. Qian, "An energy efficient virtual machine placement algorithm with balanced resource utilization," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2013 Seventh International Conference on*. IEEE, 2013, pp. 313–319.
- [32] Z. Huang and D. H. Tsang, "Sla guaranteed virtual machine consolidation for computing clouds," in *Communications (ICC), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1314–1319.
- [33] Z. Huang, D. H. Tsang, and J. She, "A virtual machine consolidation framework for mapreduce enabled computing clouds," in *Proceedings of the 24th International Teletraffic Congress*. International Teletraffic Congress, 2012, p. 26.
- [34] I. Hwang and M. Pedram, "Hierarchical virtual machine consolidation in a cloud computing system," in *Cloud Computing (CLOUD), 2013 IEEE Sixth International Conference on*. IEEE, 2013, pp. 196–203.
- [35] J. W. Jiang, T. Lan, S. Ha, M. Chen, and M. Chiang, "Joint vm placement and routing for data center traffic engineering," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 2876–2880.
- [36] H. Jin, H. Qin, S. Wu, and X. Guo, "Ccap: A cache contention-aware virtual machine placement approach for hpc cloud," *International Journal of Parallel Programming*, pp. 1–18, 2013.
- [37] H. Jin, D. Pan, J. Xu, and N. Pissinou, "Efficient vm placement with multiple deterministic and stochastic resources in data centers," in *Global Communications Conference (GLOBECOM), 2012 IEEE*. IEEE, 2012, pp. 2505–2510.
- [38] D. Kakadia, N. Kopri, and V. Varma, "Network-aware virtual machine consolidation for large data centers," in *Proceedings of the Third International Workshop on Network-Aware Data Management*. ACM, 2013, p. 6.
- [39] B. Kantarci, L. Foschini, A. Corradi, and H. T. Mouftah, "Inter-and-intra data center vm-placement for energy-efficient large-scale cloud

- systems,” in *Globecom Workshops (GC Wkshps), 2012 IEEE*. IEEE, 2012, pp. 708–713.
- [40] N. Kord and H. Haghighi, “An energy-efficient approach for virtual machine placement in cloud based data centers,” in *Information and Knowledge Technology (IKT), 2013 5th Conference on*. IEEE, 2013, pp. 44–49.
- [41] K. Le, R. Bianchini, J. Zhang, Y. Jaluria, J. Meng, and T. D. Nguyen, “Reducing electricity cost through virtual machine placement in high performance computing clouds,” in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 2011, p. 22.
- [42] K. Li, J. Wu, and A. Blaisse, “Elasticity-aware virtual machine placement for cloud datacenters,” in *Cloud Networking (CloudNet), 2013 IEEE 2nd International Conference on*. IEEE, 2013, pp. 99–107.
- [43] K. Li, H. Zheng, and J. Wu, “Migration-based virtual machine placement in cloud systems,” in *Cloud Networking (CloudNet), 2013 IEEE 2nd International Conference on*. IEEE, 2013, pp. 83–90.
- [44] W. Li, J. Tordsson, and E. Elmroth, “Virtual machine placement for predictable and time-constrained peak loads,” in *Economics of Grids, Clouds, Systems, and Services*. Springer, 2012, pp. 120–134.
- [45] X. Li, Z. Qian, S. Lu, and J. Wu, “Energy efficient virtual machine placement algorithm with balanced and improved resource utilization in a data center,” *Mathematical and Computer Modelling*, vol. 58, no. 5, pp. 1222–1235, 2013.
- [46] J.-W. Lin and C.-H. Chen, “Interference-aware virtual machine placement in cloud computing systems,” in *Computer & Information Science (ICCIS), 2012 International Conference on*, vol. 2. IEEE, 2012, pp. 598–603.
- [47] F. López-Pires and B. Barán, “A many-objective optimization framework for virtualized datacenters,” in *Proceedings of the 2015 5th International Conference on Cloud Computing and Service Science*, 2015, pp. 439–450.
- [48] F. López-Pires and B. Barán, “Virtual machine placement literature review,” Polytechnic School, National University of Asunción, Tech. Rep., 2015. [Online]. Available: <http://arxiv.org/abs/1506.01509>
- [49] F. López-Pires and B. Barán, “A virtual machine placement taxonomy,” in *Proceedings of the 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud and Grid Computing*. IEEE Computer Society, 2015.
- [50] F. López-Pires, E. Melgarejo, and B. Barán, “Virtual machine placement: a multi-objective approach,” in *Computing Conference (CLEI), 2013 XXXIX Latin American*. IEEE, 2013, pp. 1–8.
- [51] K. Lu, R. Yahyapour, P. Wieder, C. Kotsokalis, E. Yaqub, and A. I. Jehangiri, “Qos-aware vm placement in multi-domain service level agreements scenarios,” in *Cloud Computing (CLOUD), 2013 IEEE Sixth International Conference on*. IEEE, 2013, pp. 661–668.
- [52] K. Mills, J. Filliben, and C. Dabrowski, “Comparing vm-placement algorithms for on-demand clouds,” in *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*. IEEE, 2011, pp. 91–98.
- [53] M. Mishra and A. Sahoo, “On theory of vm placement: Anomalies in existing methodologies and their mitigation using a novel vector based approach,” in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*. IEEE, 2011, pp. 275–282.
- [54] I. S. Moreno, R. Yang, J. Xu, and T. Wo, “Improved energy-efficiency in cloud datacenters with interference-aware virtual machine placement,” in *Autonomous Decentralized Systems (ISADS), 2013 IEEE Eleventh International Symposium on*. IEEE, 2013, pp. 1–8.
- [55] J. Ortigoza, F. López-Pires, and B. Barán, “Workload trace generation for dynamic environments in cloud computing,” Polytechnic School, National University of Asunción, Tech. Rep., 2015. [Online]. Available: <http://arxiv.org/abs/1507.00090>
- [56] J. T. Piao and J. Yan, “A network-aware virtual machine placement and migration approach in cloud computing,” in *Grid and Cooperative Computing (GCC), 2010 9th International Conference on*. IEEE, 2010, pp. 87–92.
- [57] J. J. Prevost, K. Nagothu, B. Kelley, and M. Jamshidi, “Optimal update frequency model for physical machine state change and virtual machine placement in the cloud,” in *System of Systems Engineering (SoSE), 2013 8th International Conference on*. IEEE, 2013, pp. 159–164.
- [58] B. C. Ribas, R. M. Suguimoto, R. A. Montano, F. Silva, and M. Castilho, “Pbvmc: A new pseudo-boolean formulation to virtual-machine consolidation,” in *Intelligent Systems (BRACIS), 2013 Brazilian Conference on*. IEEE, 2013, pp. 201–206.
- [59] B. C. Ribas, R. M. Suguimoto, R. A. Montano, F. Silva, L. de Bona, and M. A. Castilho, “On modelling virtual machine consolidation to pseudo-boolean constraints,” in *Advances in Artificial Intelligence-IBERAMIA 2012*. Springer, 2012, pp. 361–370.
- [60] L. Salimian and F. Safi, “Survey of energy efficient data centers in cloud computing,” in *Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing*. IEEE Computer Society, 2013, pp. 369–374.
- [61] K. Sato, M. Samejima, and N. Komoda, “Dynamic optimization of virtual machine placement by resource usage prediction,” in *Industrial Informatics (INDIN), 2013 11th IEEE International Conference on*. IEEE, 2013, pp. 86–91.
- [62] L. Shi, B. Butler, D. Botvich, and B. Jennings, “Provisioning of requests for virtual machine sets with placement constraints in iaas clouds,” in *Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on*. IEEE, 2013, pp. 499–505.
- [63] W. Shi and B. Hong, “Towards profitable virtual machine placement in the data center,” in *Utility and Cloud Computing (UCC), 2011 Fourth IEEE International Conference on*. IEEE, 2011, pp. 138–145.
- [64] S. Shigeta, H. Yamashima, T. Doi, T. Kawai, and K. Fukui, “Design and implementation of a multi-objective optimization mechanism for virtual machine placement in cloud computing data center,” in *Cloud Computing*. Springer, 2013, pp. 21–31.
- [65] N. A. Singh and M. Hemalatha, “Reduce energy consumption through virtual machine placement in cloud data centre,” in *Mining Intelligence and Knowledge Exploration*. Springer, 2013, pp. 466–474.
- [66] F. Song, D. Huang, H. Zhou, H. Zhang, and I. You, “An optimization-based scheme for efficient virtual machine placement,” *International Journal of Parallel Programming*, vol. 42, no. 5, pp. 853–872, 2014.
- [67] V. Soundararajan and K. Govil, “Challenges in building scalable virtualized datacenter management,” *ACM SIGOPS Operating Systems Review*, vol. 44, no. 4, pp. 95–102, 2010.
- [68] B. Speitkamp and M. Bichler, “A mathematical programming approach for server consolidation problems in virtualized data centers,” *Services Computing, IEEE Transactions on*, vol. 3, no. 4, pp. 266–278, 2010.
- [69] M.-H. Tsai, J. Chou, and J. Chen, “Prevent vm migration in virtualized clusters via deadline driven placement policy,” in *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*, vol. 1. IEEE, 2013, pp. 599–606.
- [70] K. Tsakalozos, M. Roussopoulos, and A. Delis, “Vm placement in non-homogeneous iaas-clouds,” in *Service-Oriented Computing*. Springer, 2011, pp. 172–187.
- [71] S. Wang, Z. Liu, Z. Zheng, Q. Sun, and F. Yang, “Particle swarm optimization for energy-aware virtual machine placement optimization in virtualized data centers,” in *Parallel and Distributed Systems (ICPADS), 2013 International Conference on*. IEEE, 2013, pp. 102–109.
- [72] S.-H. Wang, P. P.-W. Huang, C. H.-P. Wen, and L.-C. Wang, “Eqvmp: Energy-efficient and qos-aware virtual machine placement for software defined datacenter networks,” in *Information Networking (ICOIN), 2014 International Conference on*. IEEE, 2014, pp. 220–225.
- [73] W. Wang, H. Chen, and X. Chen, “An availability-aware virtual machine placement approach for dynamic scaling of cloud applications,” in *Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2012 9th International Conference on*. IEEE, 2012, pp. 509–516.
- [74] J.-J. Wu, P. Liu, and J.-S. Yang, “Workload characteristics-aware virtual machine consolidation algorithms,” in *Proceedings of the 2012 IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE Computer Society, 2012, pp. 42–49.
- [75] K. Zamanifar, N. Nasri, and M. Nadimi-Shahraki, “Data-aware virtual machine placement and rate allocation in cloud environment,” in *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on*. IEEE, 2012, pp. 357–360.
- [76] X. Zhang, Y. Zhang, X. Chen, K. Liu, G. Huang, and J. Zhan, “A relationship-based vm placement framework of cloud environment,” in *Proceedings of the 2013 IEEE 37th Annual Computer Software and Applications Conference*. IEEE Computer Society, 2013, pp. 124–133.
- [77] X. Zhang, Q. Yue, and Z. He, “Dynamic energy-efficient virtual machine placement optimization for virtualized clouds,” in *Proceedings of the 2013 International Conference on Electrical and Information Technologies for Rail Transportation (EITRT2013)-Volume II*. Springer, 2014, pp. 439–448.